


[Home](#) [Your Room](#) [Login](#) [Contact](#) [Feedback](#)
About Us

Overview
Getting here
Committees

Products

Forecasts
Order Data
Order Software

Services

Computing
Archive
PrepIFS

Research

Modelling
Reanalysis
Seasonal

Publica

Newslett
Manuals
Library

[Home](#) > [Newsevents](#) > [Training](#) > [Rcourse_notes](#) > [DATA ASSIMILATION](#) > [ASSIM_CONCEPTS](#) >

Data assimilation concepts and me

March 1999

By F. Bouttier and P. Courtier

Table of contents

1. Basic concepts in data assimilation

2. The state vector, control space and observations

3. The modelling of errors

4. Statistical interpolation with least-squares estimation

5. A simple scalar illustration of least-squares estimation

6. Models of error covariance

7. Optimal interpolation (OI) analysis

8. Three-dimensional variational analysis (3D-Var)

9. 1D-Var and other variational analysis systems

Training Course Notes Front Page >>

Table of contents >>

Next Section >>

Previous Section >>

4 . Statistical interpolation with least-squares €

In this section we present the fundamental equation for linear analysis form: the *least squares estimation*, also called *Best Linear Unbiased*. Following sections will provide more explanations and illustrations, least-squares estimation can be simplified to yield the most common nowadays in meteorology and oceanography.

4.1 Notation and hypotheses

The dimension of the model state is n and the dimension of the observations will denote:

\mathbf{x}_t true model state (dimension n)

\mathbf{x}_b background model state (dimension n)

\mathbf{x}_a analysis model state (dimension n)

\mathbf{y} vector of observations (dimension p)

H observation operator (from dimension n to p)

B covariance matrix of the background errors ($\mathbf{x}_b - \mathbf{x}_t$) (dimension $n \times n$)

10. Four-dimensional variational assimilation (4D-Var)

11. Estimating the quality of the analysis

12. Implementation techniques

13. Dual formulation of 3D/4D-Var (PSAS)

14. The extended Kalman filter (EKF)

15. Conclusion

Appendix A. A primer on linear matrix algebra

Appendix B. Practical adjoint coding

Appendix C. Exercises

Appendix D. Main symbols

References

R covariance matrix of observation errors ($y - H[x_t]$) (dimension $n \times n$)

A covariance matrix of the analysis errors ($x_a - x_t$) (dimension $n \times n$)

The following hypotheses are assumed:

- **Linearized observation operator**: the variations of the observation operator in the vicinity of the background state are linear: for any x close enough to x_b , $H(x) - H(x_b) = H(x - x_b)$ where **H** is a linear operator.
- **Non-trivial errors**: **B** and **R** are positive definite matrices.
- **Unbiased errors**: the expectation of the background and observation errors is zero, i.e. $\overline{x_b - x_t} = \overline{y - H(x_t)} = 0$
- **Uncorrelated errors**: observation and background errors are uncorrelated, i.e. $(x_b - x_t)(y - H[x_t])^T = 0$
- **Linear analysis**: we look for an analysis defined by corrections which depend linearly on background observation departures
- **Optimal analysis**: we look for an analysis state which is as close as possible to the true state in an r.m.s. sense (i.e. it is a minimum variance estimate)

ref: Daley 1991; Lorenc 1986; Ghil 1989

4.2 Theorem: least-squares analysis equations

(a) The *optimal least-squares estimator*, or *BLUE analysis*, is defined by the following interpolation equations:

$$x_a = x_b + K(y - H[x_b])$$

$$K = B H^T (H B H^T + R)^{-1}$$

where the linear operator **K** is called the *gain*, or *weight matrix*.

(a) The *analysis error covariance matrix* is, for any **K** :

$$A = (I - K H) B (I - K H)^T + K R K^T$$

If **K** is the optimal least-squares gain, the expression becomes

$$A = (I - K H) B$$

(a) The BLUE analysis is equivalently obtained as a solution to an *optimization problem*:

$$\begin{aligned}\mathbf{x}_a &= \text{Arg min } J \\ J(\mathbf{x}) &= (\mathbf{x} - \mathbf{x}_b)^T \mathbf{B}^{-1} (\mathbf{x} - \mathbf{x}_b) + (\mathbf{y} - H[\mathbf{x}])^T \mathbf{R}^{-1} (\mathbf{y} - H[\mathbf{x}]) \\ &= J_b(\mathbf{x}) + J_o(\mathbf{x})\end{aligned}$$

where J is called the *cost function* of the analysis (or *misfit*), J_b is the *background term*, J_o is the *observation term*.

(a) The analysis \mathbf{x}_a is *optimal*: it is closest in an r.m.s. sense

(b) If the background and observation error pdfs are Gauss, \mathbf{x}_a is the *maximum likelihood estimator* of \mathbf{x}_t .

Proof:

With a translation of \mathbf{x} by \mathbf{x}_b , we can assume that $H = \mathbf{H} \mathbf{x}$ is a linear operator for our purposes. The equation (A1) is simply an expression of the fact that we want the analysis to depend linearly on the departures. The expression of \mathbf{K} in (A2) is well-defined because \mathbf{K} is a matrix, and $\mathbf{H} \mathbf{B} \mathbf{H}^T$ is positive definite. The minimization problem (A5) is a convex function and J_b is a strictly convex function (it

The equivalence between items (a) and (c) of the theorem is simply a requirement that the gradient of J is zero at the optimum \mathbf{x}_a :

$$\begin{aligned}\nabla J(\mathbf{x}_a) &= 0 = 2\mathbf{B}^{-1}(\mathbf{x}_a - \mathbf{x}_b) - 2\mathbf{H}^T \mathbf{R}^{-1}(\mathbf{y} - H[\mathbf{x}_a]) \\ 0 &= \mathbf{B}^{-1}(\mathbf{x}_a - \mathbf{x}_b) - \mathbf{H}^T \mathbf{R}^{-1}(\mathbf{y} - H[\mathbf{x}_a]) - \mathbf{H}^T \mathbf{R}^{-1} H(\mathbf{x}_a - \mathbf{x}_b) \\ (\mathbf{x}_a - \mathbf{x}_b) &= (\mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{R}^{-1}(\mathbf{y} - H[\mathbf{x}_b])\end{aligned}$$

The identity with (A2) is straightforward to prove (all inverse operators are positive definite):

$$\begin{aligned}\mathbf{H}^T \mathbf{R}^{-1} (\mathbf{H} \mathbf{B} \mathbf{H}^T + \mathbf{R}) &= (\mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H}) \mathbf{B} \mathbf{H}^T \\ &= \mathbf{H}^T + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} \mathbf{B} \mathbf{H}^T\end{aligned}$$

hence

$$(\mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{R}^{-1} = \mathbf{B} \mathbf{H}^T (\mathbf{H} \mathbf{B} \mathbf{H}^T + \mathbf{R})^{-1}$$

The expressions (A3) and (A4) for \mathbf{A} are obtained by rewriting (A1) in terms of the background, analysis and observation error

$$\begin{aligned}\varepsilon_b &= \mathbf{x}_b - \mathbf{x}_t \\ \varepsilon_a &= \mathbf{x}_a - \mathbf{x}_t \\ \varepsilon_o &= \mathbf{y} - H[\mathbf{x}_t] \\ \varepsilon_a - \varepsilon_b &= \mathbf{K}(\varepsilon_o - \mathbf{H}\varepsilon_b) \\ \varepsilon_a &= (\mathbf{I} - \mathbf{K}\mathbf{H})\varepsilon_b + \mathbf{K}\varepsilon_o\end{aligned}$$

By developing the expression of $\varepsilon_a \varepsilon_a^T$ and taking its expectation operator one finds the general expression (A3) (remembering that the errors are uncorrelated, their cross-covariance is zero). The simple derivation is obtained by substituting the expression for the optimal \mathbf{K} and simplifying.

Finally to prove (A2) itself we take the analysis error covariance and we minimize its trace, i.e. the total error variance: (note that

$$\text{Tr}(\mathbf{A}) = \text{Tr}(\mathbf{B}) + \text{Tr}(\mathbf{K} \mathbf{H} \mathbf{B} \mathbf{H}^T \mathbf{K}^T) - 2 \text{Tr}(\mathbf{B} \mathbf{H}^T \mathbf{K}^T) + \text{Tr}(\mathbf{K} \mathbf{R} \mathbf{K}^T)$$

This is a continuous differentiable scalar function of the coefficient \mathbf{K} . We express its derivative $d_{\mathbf{K}}$ as the first-order terms in \mathbf{K} of the difference $\text{Tr}(\mathbf{A})(\mathbf{K} + \mathbf{L}) - \text{Tr}(\mathbf{A})(\mathbf{K})$, \mathbf{L} being an arbitrary test matrix:

$$\begin{aligned}
d_{\mathbf{K}}[\text{Tr}(\mathbf{A})]\mathbf{L} &= 2\text{Tr}(\mathbf{K}\mathbf{H}\mathbf{B}\mathbf{H}^T\mathbf{L}^T) - 2\text{Tr}(\mathbf{B}\mathbf{H}^T\mathbf{L}^T) + 2\text{Tr}(\mathbf{K}\mathbf{R}) \\
&= 2\text{Tr}(\mathbf{K}\mathbf{H}\mathbf{B}\mathbf{H}^T\mathbf{L}^T - \mathbf{B}\mathbf{H}^T\mathbf{L}^T + \mathbf{K}\mathbf{R}) \\
&= 2\text{Tr}\{[\mathbf{K}(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R}) - \mathbf{B}\mathbf{H}^T]\mathbf{L}^T\}
\end{aligned}$$

The last line shows that the derivative is zero for any choice $(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})\mathbf{K}^T - \mathbf{B}\mathbf{H}^T = 0$, which is equivalent to

$$\mathbf{K} = \mathbf{B}\mathbf{H}^T(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1}$$

because $(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})$ is assumed to be invertible.

In the case of Gaussian pdfs, one can model the background, observation and analysis error pdfs as follows, respectively: (b , o and a are normalization factors.)

$$\begin{aligned}
\mathcal{P}_b(\mathbf{x}) &= b \exp\left[-\frac{1}{2}(\mathbf{x} - \mathbf{x}_b)^T \mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}_b)\right] \\
\mathcal{P}_o(\mathbf{x}) &= o \exp\left[-\frac{1}{2}(\mathbf{y} - \mathbf{H}[\mathbf{x}])^T \mathbf{R}^{-1}(\mathbf{y} - \mathbf{H}[\mathbf{x}])\right] \\
\mathcal{P}_a(\mathbf{x}) &= \mathcal{P}_b(\mathbf{x})\mathcal{P}_o(\mathbf{x})
\end{aligned}$$

which yields the right averages and covariances for the background and the analysis error pdf is simply defined as the Bayesian product of the two sources of information, the background and the observation pdfs (this can be justified rigorously by using Bayes' theorem to write \mathcal{P}_a as a conditional probability for observations and the a priori pdf of the background). Then, by taking the derivative of $\mathcal{P}_a(\mathbf{x})$, one finds that the model state with the maximum probability that minimizes the cost function $J(\mathbf{x})$ expressed in the theorem.

4.3 Comments

The hypotheses of non-triviality can always be made in well-posed problems. If the analysis is non-positive, one can restrict the control space to the orthogonal complement of the null space of \mathbf{H} . If the analysis will not make any correction to background variables that are not observed, then some observations are redundant and the analysis can be restricted to the image of \mathbf{H} . If \mathbf{R} is non-positive, the expression for the analysis error covariance (then the analysis will be equal to the observed value at the observation points).

the variational version of the least-squares analysis cannot be use some algebraic precautions) to have some infinite eigenvalues in J which means that some observations are not used because their e

The hypothesis of *unbiased errors* is a difficult one in practice because significant biases in the background fields (caused by biases in the the observations (or in the observation operators). If the biases are subtracted from the background and observation values, and the a the debiased quantities. If the biases are left in, the analysis will not it will seem to reduce the biases by interpolating between the background. It is important to monitor the biases in an assimilation system, e.g. background departures, but it is not trivial to decide which part of the observation biases. The problem of bias monitoring and removal is research.

The hypothesis of *uncorrelated errors* is usually justified because the background and in the observations are supposed to be completely one must be careful about observation preprocessing practices (such procedures) that use the background field in a way that biases the background. It might reduce the apparent background departures, analysis to be suboptimal (too close to the background, a condition *problem*).

The *tangent linear hypothesis* is not trivial and it is commented in the

It is possible to rewrite the least-squares analysis equations in terms of error covariance matrices, called *information* matrices. It makes the complicated, but it allows one to see clearly that the *information* is the sum, in a simple sense, of the observations provided by the background observations. This is illustrated in the section on the estimation of a

It will be shown in the section on dual algorithms (PSAS analysis) that particular the cost function J , can be rewritten in the space of the background easy to that least-squares analysis is closely related to a linear regression state and observations.

4.4 On the tangent linear hypothesis

The hypothesis of *linearized observation operator* is needed in order to have an algebraic expression for the optimal K . In practice, H may not be linear, but it makes physical sense to linearize it in the vicinity of the background

$$H(\mathbf{x}) - H(\mathbf{x}_b) \approx \mathbf{H}(\mathbf{x} - \mathbf{x}_b)$$

Then, \mathbf{K} being a continuous function of \mathbf{H} , the least-squares equations should intuitively yield a nearly optimal \mathbf{x}_a .

More generally, the *tangent linear hypothesis* on H can be written Young formula in the vicinity of an arbitrary state \mathbf{x} and for a perturbation h

$$H(\mathbf{x} + h) = H(\mathbf{x}) + \mathbf{H}h + O(\|h\|^2)$$

with $\lim_{h \rightarrow 0} O(\|h\|^2)h^{-2} = 0$. This hypothesis, called the *tangent linear hypothesis*, is acceptable if the higher-order variations of H can be neglected (in the absence of discontinuities) for all perturbations of the model state which have a magnitude as the background errors. The operator \mathbf{H} is called the *derivative*, or *tangent linear (TL)*¹ function of H at point \mathbf{x} . Although a mathematical property of H , it is not enough for practical purposes to approximate

$$H(\mathbf{x} + h) - H(\mathbf{x}) \approx \mathbf{H}h$$

must be satisfactory, in user-defined terms, for finite values of h in the application considered. In the least-squares analysis problem, we need

$$\mathbf{y} - H(\mathbf{x}) \approx \mathbf{y} - H(\mathbf{x}_b) + H(\mathbf{x}_b)$$

for all values of \mathbf{x} that will be encountered in the analysis procedure and also all trial values used in the minimization of $J(\mathbf{x})$ if a variational method is performed². Thus the important requirement is that the difference between \mathbf{x} and \mathbf{x}_b should be much smaller than the typical observation error. Model state perturbations $\mathbf{x} - \mathbf{x}_b$ of size and structure consistent with the background errors, and also with the amplitude of the analysis increments $\mathbf{x}_a - \mathbf{x}_b$:

Thus the problem of linearizing H is not just related to the observation errors. It must be appreciated in terms of the errors in the background \mathbf{x}_b too. In a data assimilation system are the previous forecast errors, which depend on the model and the quality of the model. Ultimately the correctness of the linearization must be appreciated in the context of the fully integrated assimilation system. The linearization is applied to a good system because the departures $\mathbf{x} - \mathbf{x}_b$ within the linearization may be inapplicable to difficult data assimilation problems, such as with ocean models or satellite data, which means that it can be improved by sophisticated analysis algorithms that rely too much on the linearity

The linearization problem can be even more acute for the linearization operator M which is needed in 4D-Var and in the Kalman filter. In the linearization of H , it may or may not be licit depending on the quality of the assimilation system: data coverage, observation quality, model and forecast range. The user requirements and the physical properties must be considered.

The non-linear analysis problem

The assumption of linear analysis is a strong one. Linear algebra is optimal for analysis equations. One can rely on the linearization of a weak observation operator, at the expense of optimality. The incremental variational analysis performs this procedure iteratively in an attempt to make the analysis more optimal. For strongly non-linear problems, there is no simple way to calculate the optimal analysis. The simulated annealing and specific methods, such as the simplex, deal with variables with boundaries. Finally, it is sometimes possible to make a problem more linear by separating model and observation variables (see the section on minimization).

4.5 The point of view of conditional probabilities

It is interesting to formalize the analysis problem using the conditional probabilities. Let us denote $P(x)$ the a priori pdf (probability density function) of the model state before the observations are considered, i.e. the background pdf. Let $P(y)$ be the pdf of the observations. The aim of the analysis is to find the maximum a posteriori conditional probability of the model state given the observations. The conditional probability $P(x|y)$ (i.e. the probability that x and y occur together) is

$$P(x \wedge y) = P(x|y)P(y) = P(y|x)P(x)$$

i.e. it is the probability that x occurs when y occurs, and vice versa according to the Bayes theorem. In the analysis procedure we know that a measurement y has been made and we know its value y , so $P(y) = 1$ and we obtain

$$P(x|y) = P(y|x)P(x)$$

which means that the analysis pdf is equal to the background pdf $P(x)$ multiplied by $P(y|x)$. The latter peaks at $y = H(x)$ but it is not a Dirac distribution because observations are not error-free.

The virtue of the probabilistic derivation of the analysis problem is that it handles non-Gaussian probabilities (although this spoils the equivalence with the Kalman filter). A practical application is done in the framework of variational data assimilation.

assumed that observation errors are not Gaussian but they contain errors", i.e. there is a probability that the error is not generated by physical processes but by some more serious problem, like coding. The gross errors might be modelled using a uniform pdf over a pre-admissible gross errors, leading to a non-Gaussian observation pdf. If the logarithm of this pdf is taken, the resulting observation cost function gives less weight to the observation (i.e. there is less slope) for more strongly with the observed value.

ref: Lorenc 1986

4.6 Numerical cost of least-squares analysis

In current operational meteorological models, the dimension of the precisely, of the control variable space) \mathbf{x} is of the order of $n = 10^6$. The observation vector (the number of observed scalars) is of the order of $p = 10^3$. Therefore the analysis problem is mathematically underdetermined in some regions it might be overdetermined if the density of the observations is higher than the resolution of the model). In any practical application it is essential to use efficient matrix operators involved in computing the analysis (Fig. 4). The method requires in principle the specification of covariance matrices (the analysis and the observation) which respectively have $n^2/2$ and $p^2/2$ distinct coefficients, which are statistics to estimate variance or covariance. The iterative method converges like the square root of the number of iterations. The explicit determination of \mathbf{K} requires the inversion of a matrix of size $n \times n$. The asymptotic complexity of the order of $p^2 \log(n)$. The exact minimization of J requires, in principle, $n + 1$ evaluations of the cost function and is quadratic and there are no numerical errors (e.g. using a conjugate

$$x_a = x_b + K(y - Hx_b)$$

$$K = BH^T (HBH^T + R)^{-1}$$

$$HBH^T$$

$$J(x) = (x - x_b)^T B^{-1} (x - x_b) + (y - Hx)^T R^{-1} (y - Hx)$$

Figure 4 . Sketches of the shapes of the matrices and vector d in a usual analysis problem where there are many fewer observations than degrees of freedom in the model: from top to bottom, in the equations of computation of K , of the HBH^T term, and the computation of

It is obvious that, except in analysis problems of very small dimensions (retrievals), it is impossible to compute exactly the least-squares analysis. If approximations are necessary, they are the subject of the following

4.7 Conclusion

We have seen that there are two main ways of defining the statistic

- either assume that the background and error covariances define the analysis equations by requiring that the total analysis error variance is a minimum
- or assume that the background and observation error pdfs define the analysis equations by looking for the state with the maximum probability

Both approaches lead to two mathematically equivalent algorithms

- the direct determination of the analysis gain matrix K ,

- the minimization of a quadratic cost function.

These algorithms have very different numerical properties, and the soon as some underlying hypotheses are not verified, like the linear operator, for instance.

[Training Course Notes Front Page >>](#)

[Table of contents >>](#)

[Next Section >>](#)

[Previous Section >>](#)

¹ Both qualifiers *tangent* and *linear* are needed: obviously \mathbf{H} could satisfying the Taylor formula. A function can also be tangent to another if the difference between them is an $\mathcal{O}(\|\mathbf{h}\|^2)$, e.g. x^2 and x^3 are tangent at $x = 0$.

² Qualitatively speaking they all belong to a neighbourhood of \mathbf{x}_b which is consistent with the \mathbf{B} and \mathbf{R} error covariances.

³ At ECMWF in winter 1998 the control variable dimension was 51 observations (per 6-hour interval) was about 150000